Diese Website verwendet Cookies von Google, um Dienste anzubieten und den Traffic zu analysieren. Deine IP-Adresse und dein User-Agent werden zusammen mit Messwerten zur Leistung und Sicherheit für Google freigegeben. So können Nutzungsstatistiken generiert, Missbrauchsfälle erkannt und behoben und die Qualität des Dienstes gewährleistet werden.

WEITERE INFORMATIONEN OK

Saturday, July 14, 2012

Pascal's scams

Beware of what I call Pascal's scams: movements or belief systems that ask you to hope for or worry about very improbable outcomes that could have very large positive or negative consequences. (The name comes of course from the infinite-reward Wager proposed by Pascal: these days the large-but-finite versions are far more pernicious). Naive expected value reasoning implies that they are worth the effort: if the odds are 1 in 1,000 that I could win $1 billion, and I am risk and time neutral, then I should expend up to nearly $1 million dollars worth of effort to gain this boon. The problems with these beliefs tend to be at least threefold, all stemming from the general uncertainty, i.e. the poor information or lack of information, from which we abstracted the low probability estimate in the first place: because in the messy real world the low probability estimate is almost always due to low or poor evidence rather than being a lottery with well-defined odds:

(1) there is usually no feasible way to distinguish between the very improbable (say, 1 in 1,000) and the extremely improbable (e.g., one in a billion). Poor evidence leads to what James Franklin calls "low-weight probabilities", which lack robustness to new evidence. When the evidence is poor, and thus robustness of probabilities is lacking, then it is likely that "a small amount of further evidence would substantially change the probability. " This new evidence is as likely to decrease the probability by a factor of X as increase it by a factor of X, and the poorer the original evidence, the greater X is. (Indeed, given the nature of human imagination and bias, it is more likely to decrease it, for reasons described below).

(2) the uncertainties about the diversity and magnitudes of possible consequences, not just their probabilities, are also likely to be extremely high. Indeed, due to the overall poor information, it's easy to overlook negative consequences and recognize only positive ones, or vice-versa. The very acts you take to make it into utopia or avoid dystopia could easily send you to dystopia or make the dystopia worse.

(3) The "unknown unknown" nature of the most uncertainty leads to unfalsifiablity: proponents of the proposition can't propose a clear experiment that would greatly lower the probability or magnitude of consequences of their proposition: or at least, such an experiment would be far too expensive to actually be run, or cannot be conducted until after the time which the believers have already decided that the long-odds bet is rational. So not only is there poor information in a Pascal scam, but in the more pernicious beliefs there is little ability to improve the information.

The biggest problem with these schemes is that, the closer to infinitesimal probability, and thus usually to infinitesimal quality or quantity of evidence, one gets, the closer to infinity the possible extreme-consequence schemes one can dream up, Once some enterprising memetic innovator dreams up a Pascal's scam, the probabilities or consequences of these possible futures can be greatly

About Me

Nick Szabo

"A premier thinker about history, law and economics, and the lessons they have for security." -- Adam Shostack, Emergent Chaos

"Szabo comes out with these essays that leave me in awe." -- Brian Dunbar

"Reading material that is eclectic, challenging and endlessly fascinating." -- Sean McGrath, Propylon

"Like most blogs worth my attention, this blog is updated only infrequently. That is because the authors of blogs worth my attention only post when they have something to say that is true, relevant and not already known by their audience. Most of the human race does not have the skill to know when an idea has these three properties. The skill is particularly rare in the fields of politics and economics, which is why this blog is such a rare and valuable thing." -- Richard Hollerith

exaggerated yet still seem plausible. "Yes, but *what if*?" the carrier of such a mind-virus incessantly demands. Furthermore, since more than a few disasters are indeed low probability events (e.g. 9/11), the plausibility and importance of dealing with such risks seems to grow in importance after they occur -- the occurrence of one improbable disaster leads to paranoia about a large number of others, and similarly for fortuitous windfalls and hopes. Humanity can dream up a near-infinity of Pascal's scams, or spend a near-infinity of time fruitlessly worrying about them or hoping for them. There are however far better ways to spend one's time -- for example in thinking about what has actually happened in the real world, rather than the vast number of things that might happen in the future but quite probably won't, or will likely cause consequences very differently than you expect.

So how should we approach low probability hypotheses with potential high value (negative or positive) outcomes? Franklin et. al. suggest that "[t]he strongly quantitative style of education in statistics, valuable as it is, can lead to a neglect of the more qualitative, logical, legal and causal perspectives needed to understand data intelligently. That is especially so in extreme risk analysis, where there is a lack of large data sets to ground solidly quantitative conclusions, and correspondingly a need to supplement the data with outside information and with argument on individual data points."

On the above quoted points I agree with Franklin, and add a more blunt suggestion: stop throwing around long odds and dreaming of big consequences as if you are onto something profound. If you can't gather the information needed to reduce the uncertainties, and if you can't suggest experiments to make the hope or worry falsifiable, stop nightmaring or daydreaming already. Also, shut up and stop trying to convince the rest of us to join you in wasting our time hoping or worrying about these fantasies. Try spending more time learning about what has actually happened in the real world. That study, too, has its uncertainties, but they are up to infinitely smaller.

Posted by Nick Szabo at 12:16 AM

40 comments:

Brian H said...

The "qualitative" analysis basically comes down to having the cojones to call "Bullshit" when you smell it.

11:04 PM

Anonymous said...

Thanks for taking a stand against the Singularity Institute.

12:41 PM

Anonymous said...

No Hansonian plastic brains in Dewar flasks for you!

-- Singularity Nazi

3:42 PM

Blog Archive

- ► 2018 (1)
- ► 2017 (3)
- ► 2016 (4)
- ► 2015 (3)
- ► 2014 (3)
- ► 2013 (3)
- ▼ 2012 (8)
  - ► October (2)
  - ► August (2)
  - ▼ July (4)
    - Three philosophical essays
    - Pascal's scams (ii)
    - Pascal's scams
    - More short takes
- ► 2011 (12)
- ► 2010 (9)
- ► 2009 (29)
- ► 2008 (55)
- ► 2007 (47)
- ► 2006 (130)
- ► 2005 (44)

nick said...

Brian H, you'll often save many people from needless nightmaring or daydreaming if you are able to give good reasons about why it's BS.

In response to the two Anonymi, I wasn't going to name names, but the rapture or doom of "the Singularity", "SkyNet", "the Matrix", and similar robot apocalypse scenarios, does seem to fit the bill, especially regarding the inability of their proponents to design experiments that would give us information about these ideas that would greatly reduce uncertainties about them. Instead, they go straight to supposed solutions (per above, as likely to lead closer to dystopia as to closer to utopia, since the uncertainties about the underlying causes are so high, falsely precise rhetoric notwithstanding).

I describe the economic and computational efficiency reasons why the odds of either very useful or very dangerous "general AI" are very long here, see both the original post and some of my comments below it.

7:12 PM

nick said...

Per Hanson's donation: a huge uncertainty, i.e. point (1) above, with the Brain Preservation Technology Prize is the supposed strong relationship between "with such fidelity that the structure of every neuronal process and every synaptic connection remains intact and traceable using today's electron microscopic (EM) imaging techniques" (a mistake right there, since these techniques image static structure, they don't trace processes), and the goal that one day subjects undergoing some such treatment will "live again." It's quite possible that the presumably requisite consciousness and memories for "liv[ing] again" require molecular or other kinds of structures different from those being preserved here.

Furthermore, preserving information about a molecular structure does *not* automatically, or perhaps even probably in the case of sophisticated assemblages of biomolecules, lead to the ability to recreate that structure in sufficient detail. It's possible that the chemical routes for doing so may prove infeasible (i.e. too expensive for any feasible technology).

What's more, per point (2), we don't know whether such capabilities even if they worked would lead to utopia or dystopia for the experimental subjects preserving their brains in hopes of future reanimation. The superhuman brain scientists of the future mainly might well look on me, even while "liv[ing] again", the way today's scientists treat their static brain slices: as a specimin to be studied. Nobody real knows whether these experimental subjects will wake up in a high-tech paradise or as hopelessly inferior subjects in a gruesome experiment they can't control.

There are probably many other sources of uncertainty that I haven't considered yet, and many they haven't considered yet either.

On point (3) I give them credit since they are trying experiments to reduce some of the uncertainties involved, even if much uncertainty would remain after the prize for preserving certain structures has been won. And the experiments don't seem very expensive. So, while not donating myself, I say let them have at it.

7:50 PM

coldsponger said...

One might summarize this as "learn what a partial derivative is and how to use it".

10:13 PM

nazgulnarsil said...

nick, WRT plastination:
My understanding is that the ability to cold boot humans without loss of function (people resuscitated after cessation of electrical activity in the brain) implies that the structure is what is important.

12:46 AM

Alexander Kruel said...

@Nick Szabo

Eliezer Yudkowsky does actually believe that a negative technological singularity due to unfriendly artificial general intelligence is not a low probability event. Here are two quotes:

"I don't think the odds of us being wiped out by badly done AI are small. I think they're easily larger than 10%. And if you can carry a qualitative argument that the probability is under, say, 1%, then that means AI is probably the wrong use of marginal resources – not because global warming is more important, of course, but because other ignored existential risks like nanotech would be more important. I am not trying to play burden-of-proof tennis. If the chances are under 1%, that's low enough, we'll drop the AI business from consideration until everything more realistic has been handled. We could try to carry the argument otherwise, but I do quite agree that it would be a nitwit thing to do in real life, like trying to shut down the Large Hadron Collider."

http://johncarlosbaez.wordpress.com/2011/04/24/what-to-do/#comment-5546

"I keep telling SIAI people to never reply to "AI risks are small" with "but a small probability is still worth addressing". Reason 2: It sounds like an attempt to shut down debate over probability. Reason 3: It sounds like the sort of thing people are more likely to say when defending a weak argument than a strong one. Listeners may instinctively recognize that as well.

Existential risks from AI are not under 5%. If anyone claims they are, that is, in emotional practice, an instant-win knockdown argument unless countered; it should be countered directly and aggressively, not weakly deflected."

The same seem to be true for cryonics, if you Google "We Agree: Get Froze" you can find the following quote by Robin Hanson:

" So if whole brain emulation is ever achieved, and if freezing doesn't destroy info needed for an em scan, ifs we think more likely than not, future folks could make an em out of your frozen brain."

2:14 AM

Carl said...

Robin on cryonics:

"Your chance of being usefully revived in 2090 as an em is roughly the product of these ten conditional probability terms. Ten 90% terms gives a total chance of ~1/3. Ten 80% terms gives a total chance of ~10%, except step 4 might be a 50% chance, for a total chance of ~6%, which seems about right to me."
http://www.overcomingbias.com/2009/03/break-cryonics-down.html

2:28 AM

Alexander Kruel said...

I don't know if you read lesswrong but some of them seem to use the fact that the Singularity Institute does believe that it is *not* improbable (>1%) that we'll have to face AI risks as an argumentative defeat.

Which doesn't really undermine your argument in my opinion. Since as far as I understand, you 1.) do not believe that a negative

singularity is probable 2.) don't think that the burden of proof is on you to show that it isn't probable 3.) believe that the margin of error of any argument to the contrary is large enough to constitute a case of Pascal's mugging, since claims by the mugger that the threat is probable are simply part of the scenarios that you reject.

I have written down some of my on throughts on the topic.

7:29 AM

nick said...

Alexander: that's largely correct. Regarding your point (2), I do think it's useful for counter-arguments to exist to make at least a plausible case for the low probability of the general category of the futuristic scenario, as I have regarding specialized vs. general AI. One is under no obligation at all to address their specific claims, much less their specific solutions, as there are a near-infinity of such possible claims and solutions.

This is the first time I've heard the phrase "Pascal's mugging", so I just Googled it and saw Yudkowsky's article -- a similar idea to the Pascal scam to be sure, but my focus is on the underlying poor quality of evidence for the claim, in situations that don't need to be anywhere close to as theoretical or extreme as the idea that we are living in a simulation. There are a near-infinity of plausible-sounding but poorly evidenced claims about the future one can make in and about our own physical universe, so that such claims should be nearly infinitely discounted.

The Pascal's mugger, who implicitly states that he is 100% certain, is a good example of why numerical probability estimates in these situations are false precision and arguments from authority, not good evidence. Some people who call themselves Bayesians sometimes or even often tend to confuse probability estimates with actual evidence, obsessing with the estimates and ignoring the actual evidence (or lack thereof).

2:11 PM

nick said...

Also, "Pascal's mugger" strongly suggests some intent to deceive, whereas no such intent is required for a Pascal scam. (Even "scam" may thus be often less accurate terminology than, say, "delusion", or in non-obsessive and non-evangelizing cases just "mistake").

2:16 PM

nick said...

nazgulnarsil, that structure could well be molecular, not just the connectome (synapses plus their connections). That possibility is plausible enough that the burden of proof is on those making such otherwise meaningless numerical probability estimates to prove otherwise, and thus justify the attention, money, or other resources they seek.

2:26 PM

nick said...

Here's a comment I made to people responding to (at least a lossy summary of) this post on Less Wrong:

The asteroid threat is a good example of a low-probability disaster that is probably not a Pascal scam. On point (1) it is fairly lottery-like, insofar as asteroid orbits are relatively predictable -- the unknowns are primarily "known unknowns", being deviations from very simple functions -- so it's possible to compute odds from actual data, rather than merely guessing them from a morass of "unknown unknowns". It passes test (2) as we have good ways to simulate with reasonable accuracy and (at some expense, only if needed) actually test solutions. And best of all it passes test (3) -- experiments or observations can be done to improve our information about those odds. Most of the funding has, quite properly, gone to those empirical observations, not towards speculating about solutions

before the problem has been well characterized.

Alas, most alleged futuristic threats and hopes don't fall into such a clean category: the evidence is hopelessly equivocal (even if declared with a false certainty) or missing, and those advocating that our attention and other resources be devoted to them usually fail to propose experiments or observations that would imrove that evidence and thus reduce our uncertainty to levels that would distinguish them from the near-infinity of plausible disaster scenarios we could imagine. (Even with just the robot apocalypse, there are a near-infinity of ways one can plausibly imagine it playing out).

2:31 PM

Alexander Kruel said...

This comment has been removed by the author.

1:47 AM

Alexander Kruel said...

Some people who call themselves Bayesians sometimes or even often tend to confuse probability estimates with actual evidence, obsessing with the estimates and ignoring the actual evidence (or lack thereof).

Which is basically the gist of most disagreement between the kind of people associated with the Singularity Institute and who those people call "traditional rationalists".

The current state of the art formalization of technical rationality does indicate that a rational decision maker should rely less on empirical evidence than "traditional rationalists"/scientists tend to do and rather take logical implications much more seriously.

My personal disagreement with them is mainly that I go a step further in rejecting the extent to which logical implications bear on decision making. Just as them I wouldn't give money to a Pascal's mugger. But I consider certain attitudes towards AI risks and existential risks to be a case of Pascal's mugging. Which they don't.

As far as I can tell, the whole sequence of posts that Eliezer Yudkowsky wrote on the many world interpretation of quantum mechanics was meant to convey what you are rejecting, namely that probabilistic beliefs should be taken more seriously. And further that the implied invisible, that which logically follows from any given empirical evidence, shouldn't be discounted completely.

I might be wrong here as I am not too familiar with the sequences.

I don't disagree with them, except that I think the associated model uncertainty is doomed to be too drastic to take any given conclusions too seriously. More here.

In short (excerpt of my post):

You might argue that I would endorse position 12 [1] if NASA told me that there was a 20% chance of an extinction sized asteroid hitting Earth and that they need money to deflect it. I would indeed. But that seems like a completely different scenario to me.

I would have to be able to assign more than a 80% probability to AI being an *existential* risk to endorse that position. I would further have to be *highly* confident that we will have to face that risk *within this century* and that the model uncertainty associated with my estimates is *low*.

That intuition does stem from the fact that any estimates regarding AI risks are *very* likely to be wrong, whereas in the example case of an asteroid collision one could be *much* more confident in the 20% estimate. As the latter is based on *empirical evidence* while the former is inference based and therefore much more **error prone**.

I don't think that the evidence allows anyone to take position 12, or even 11, and be even slightly confident about it.

I am also highly skeptical about using the expected value of a galactic civilization to claim otherwise. Because that reasoning will ultimately make you privilege unlikely high-utility outcomes over much more probable theories that are based on empirical evidence.

[1] Position 12: *This is crunch time. This is crunch time for the entire human species. And it's crunch time not just for us, it's crunch time for the intergalactic civilization whose existence depends on us. Everyone should contribute all but their minimal living expenses in support of the issue.* (quote by Eliezer Yudkowsky)

1:48 AM

Dan said...

With asteroid detection, there is a mechanism to determine where we have established a dangerous asteroid isn't. We can asymptotically approach 100% certainty that there is no dangerous asteroid within any given volume. You could easily administer a $10bn bounty, payable to the first person or group who identifies an asteroid of some determined size that will collide with Earth, and a $1tn bounty to the group that manages to change the trajectory. Those bounties are clearly worthwhile to issue, even at those absurd amounts. The low chance of payoff might result in few people trying to earn them, but that's a different issue.

Right now the SI has not adequately defined Friendly AI to the point where they could offer a prize for solving the problem. If they can't offer a definition robust enough for a neutral party to resolve the dispute which will occur when somebody claims to have solved the AI problem but the SI disagrees, then they haven't defined the problem well enough to post a prize.

1:55 AM

Alexander Kruel said...

One correction regarding my last comment. Position 12 is not a verbatim quote by Eliezer Yudkowsky. Here is what he actually wrote:

[...] I would be asking for more people to make as much money as possible if they're the sorts of people who can make a lot of money and can donate a substantial amount fraction, never mind all the minimal living expenses, to the Singularity Institute.

This is crunch time. This is crunch time for the entire human species. [...] and it's crunch time not just for us, it's crunch time for the intergalactic civilization whose existence depends on us. I think that if you're actually just going to sort of confront it, rationally, full-on, then you can't really justify trading off any part of that intergalactic civilization for any intrinsic thing that you could get nowadays [...]

[...] having seen that intergalactic civilization depends on us, in one sense, all you can really do is try not to think about that, and in another sense though, if you spend your whole life creating art to inspire people to fight global warming, you're taking that 'forgetting about intergalactic civilization' thing much too far."

— Video Q&A with Eliezer Yudkowsky

2:53 AM

Alexander Kruel said...

If you wonder what AI researchers think, some time ago I conducted a Q&A style interview series with a bunch of people, all except one actual AI researchers:

http://wiki.lesswrong.com/wiki/Interview_series_on_risks_from_AI

@Nick Szabo

I would love you to answer the questions as well :-)

1.) Assuming beneficial political and economic development and that no global catastrophe halts progress, by what year would you assign a 10%/50%/90% chance of the development of artificial intelligence that is roughly as good as humans (or better, perhaps unevenly) at science, mathematics, engineering and programming?

2.) Once we build AI that is roughly as good as humans (or better, perhaps unevenly) at science, mathematics, engineering and programming, how much more difficult will it be for humans and/or AIs to build an AI which is substantially better at those activities than humans?

3.) Do you ever expect artificial intelligence to overwhelmingly outperform humans at typical academic research, in the way that they may soon overwhelmingly outperform humans at trivia contests, or do you expect that humans will always play an important role in scientific progress?

4.) What probability do you assign to the possibility of an AI with initially roughly professional human-level competence (or better, perhaps unevenly) at general reasoning (including science, mathematics, engineering and programming) to self-modify its way up to vastly superhuman capabilities within a matter of hours/days/< 5 years?

5.) How important is it to research risks associated with artificial intelligence that is good enough at general reasoning (including science, mathematics, engineering and programming) to be capable of radical self-modification, before attempting to build one?

6.) What probability do you assign to the possibility of human extinction within 100 years as a result of AI capable of self-modification (that is not provably non-dangerous, if that is even possible)? P(human extinction by AI | AI capable of self-modification and not provably non-dangerous is created)

3:02 AM

Alexander Kruel said...

Regarding the questions of my previous comment and the probability estimates they ask for. I agree with you that such estimates can be very misleading. I don't expect numerical estimates. I'd just be curious about your general opinion with respect to those questions.

It is actually one of the topics I am most confused about, namely the pros and cons of making up numerical probability estimates.

I do believe that using Bayes' rule when faced with data from empirical experiments, or goats behind doors in a gameshow, is the way to go.

But I fear that using formal methods to evaluate informal evidence might lend your beliefs an improper veneer of respectability and in turn make them appear to be more trustworthy than your intuition. For example, using formal methods to evaluate something like AI risks might cause dramatic overconfidence.

Bayes' rule only tells us by how much we should increase our credence given certain input. But given most everyday life circumstances the input is often conditioned and evaluated by our intuition. Which means that using Bayes' rule to update on evidence does emotionally push the use of intuition onto a lower level. In other words, using Bayes' rule to update on evidence that is vague (the conclusions being highly error prone), and given probabilities that are being filled in by intuition, might simply disguise the fact that you are still using your intuition after all, while making you believe that you are not.

Even worse, using formal methods on informal evidence might introduce additional error sources.

3:49 AM

Dmytry said...

Alexander:

> The current state of the art formalization of technical rationality does indicate that a rational decision maker should rely less on empirical evidence than "traditional rationalists"/scientists tend to do and rather take logical implications much more seriously.

According to Eliezer &co, who have an obvious monetary motivation, and who do not promote logical implications, but instead something that on surface sounds vaguely reasonable but is about as effective as truth finding method as medieval scholasticism.

You do not have reason to describe it as state of the art. There has been no great success attributable to informal 'formalizations' you refer to.

While the logical implications certainly have to be followed, the scholastic implications as conjectured by Eliezer, certainly should not.

Furthermore, from visiting lesswrong you may have massively skewed view on the decision making under uncertainty. There is a method that does not rely on made up priors. If I commit to believe in a hypothesis if it passes a test that has one in a million chance of a false positive, i have at most one in a million chance of believing in an invalid hypothesis; after considering the number of hypotheses you deal with, the cost of invalid belief, and the cost of tests, an appropriate testing strategy can be devised with a well defined worst-case performance, without ever making up a prior out of thin air.

The 'technical rationality' as described on lesswrong is as much worthy of 'state of the art' title as Hubbard's dianetics or Keith Raniere's NXIVM.

7:44 AM

Dmytry said...

Alexander:

> The current state of the art formalization of technical rationality does indicate that a rational decision maker should rely less on empirical evidence than "traditional rationalists"/scientists tend to do and rather take logical implications much more seriously.

What are you calling "state of the art" ? The informal "formalizations" originating from the same Eliezer Yudkowsky? What he is relying on has nothing to do with logic and as a truth finding method ranks some place near medieval scholasticism. Half the arguments rely on 'look theres this possible consequence and we can't imagine other possibilities so it must be true', and the other half replackages the general difficulty of building and motivating AI as the difficulty of making survivable AI.

11:58 AM

Alexander Kruel said...

What are you calling "state of the art" ? The informal "formalizations" originating from the same Eliezer Yudkowsky?

Bayes' Theorem, the expected utility formula, and Solomonoff induction.

12:14 PM

Dmytry said...

Alexander:

And you learnt of those from where?

I'm using Bayes theorem in my work. Can't say about Solomonoff induction, it being non-computable.

The Bayes theorem, combined with Solomonoff induction and halting problem/undecidability, is if anything, a proof that you can not find a probability of theory being true.

Instead, alternative methods have to be used when no prior is available.

For instance, if I commit to believe that a hypothesis is true if an experiment with one in a billion false positive rate has confirmed that the hypothesis is true, then the probability of me ending up with a belief in this false hypothesis is at most one in a billion. Given the cost of experiments and the costs of invalid beliefs of either kind, as well as the number of hypotheses being tested, the optimal standard for evidence (under worst-case assumptions) can be chosen, all without EVER pulling a number out of your ass and calling it a probability. This comes complete with resistance to various forms of pascal's wager, especially given that in the event that you are being offered a wager, there is significant probability of needing the money in the future for some other wager that would come with a proof.

1:27 PM

Tim said...

Another result of what I once dubbed "Rapture of the Future" is that makes it hard to work on all that boring, short-term, intense stuff....like debugging a chip, or installing a new ion implant machine, that sort of boring grunge.

I'm being facetious. But I saw it a bunch of times back 20 years ago when a lot of the cool new stuff was first getting wide exposure on the then-emerging mailing lists. (Extropians, Cryonics-l, Cyberia-l, Cypherpunks...).

A bunch of folks were doing menial work at bookstores (one worked at the libertarian bookstore in SF, another was an unemployed philosopher). It was a lot easier to think about what "rules for Jupiter-sized brains" should be than to learn, as one of your blog commenters put it, to learn what a partial derivative is and how to use it.)

I can recall one of the leader philosophers of nanotechnology--no, neither Drexler nor Merkle--saying at a nanotech discussion in 1992-3 that "This entire valley (Silicon) will be gone in 20 years!"

Whoops.

But the effect of the intoxication of the Rapture of the Future, either the positive or the negate sides of what Nick is calling Pascal's scams, can be debilitating.

Frankly, I'm pretty glad that I "came up" during an era where so much tweeting and blogging and e-mailing and "bottle rockets being fired off" were not distracting me. To learn what I needed to learn to later do something useful (and financially rewarding) I had to buckle down and work. And at my job, the focus was on fairly short-term milestones, not grand visions of the future.

Believe me, one of my boss's boss's bosses was Gordon Moore. And I can say he was no blue sky dreamer. And his observation had a lot more to do with saying "This is what we've seen in the last 5-6 years," along with some solid comments on the obvious "t-shirt printing" side of things. By this I mean that photolithography is a lot like silk-screening a t-shirt, and a lot of the jump from 100 transistors per die (chip) to 10,000 to a million, etc. had to do with precision optics, precision stepper motors....essentially printing images at higher and higher resolutions. Some things came from device physics (I worked on some of this), some things came from better CAD tools, but most of "Moore's Law" comes from photolithography. Even today, with the newest "step and repeat" systems costing $25 million EACH. Those are some expensive cameras!

Not much room for Rapture-a-tarians in this environment.

And a problem with a lot of these "planning" groups (the Usual Suspects) is that too many of the members want to be the Big Thinkers and authors of the policy papers.

--Tim May, Corralitos, CA

9:29 PM

nick said...

Tim, great comments on the actual history of computing -- far better information there than in a library full of speculations about the future. So it turns out that the future of computing depended much more on t-shirt printing tech than on self-replicating diamondoid nanobots. Put that into your Bayesian pipes and puff out some numbers. :-)

Alexander, for most of your questions, even putting an answer into words, much less numbers, would be an exercise in false precision. I don't know what the "t-shirt printing" technology of the future will be (of course the futurists who pretend to know such things don't know either, so I'm in good company). Generally speaking, though, most of those things are already partially happening now, but won't happen anywhere close to completely before culture has radically changed in a myriad of ways, by which time such questions will have become largely moot, shocking as they may seem to us today.

10:33 PM

Tim said...

Yeah, Nick, it was sort of just plain old, mundane t-shirt printing technology!

But, in its very large (economically) and very sophisticated (complexity) way, it was a blast.

I worked for Intel from 1974 to 1986. I then did other things.

But as a holder of Intel stock, and someone still interested in the technology, I follow what Intel, Samsung, TSMC, the leaders in the chip industry today, are doing.

It bears very little resemblance to the Thinking Institutes which pontificate about whether step-and-repeat cameras are Dangers to Our Existence or What Laws do We Need to Head Off the AI Takeover.

Good to see you blogging again. I would blog, maybe, except I did this for about 50% of my waking hours from 1992 to around 2002. Then I lessened my mailing list activity and got back to some more core stuff.

(Certainly the "your great thoughts in 40 characters or less" current situation is really a disaster.)

--Tim May, Corrallitos, CA

11:01 PM

Alexander Kruel said...

@Nick Szabo

Do you believe that if, at some point in future, we combine our experts systems, tools, into a coherent framework of agency, this will be of no advantage to the resulting agent that is large enough for it to pose a risk to humanity?

That's what I inferred from skimming over your conversation over at lesswrong.

12:26 PM

nick said...

Do you believe that if, at some point in future, we combine our experts systems, tools, into a coherent framework of agency, this will be of no advantage to the resulting agent that is large enough for it to pose a risk to humanity?

Stated this way, I don't even buy the "if" part -- combining software systems is a difficult engineering problem that gets much harder as the total lines of code increase. I've seen large software systems and "coherent" is not how I'd describe them. :-) Software is going to become more fragmented, not more combined.

And of course as you state I do indeed believe that to the extent such combined software is feasible it would not pose any kinds of problems that specialized software had not posed long before and to a far greater degree.

5:06 PM

Alexander Kruel said...

Wei_Dai wrote (@lesswrong):

"Given the benefits of specialization, how do you explain the existence of general intelligence (i.e. humans)? Why weren't all the evolutionary niches that humans current occupy already taken by organisms with more specialized intelligence?

My explanation is that generalized algorithms may be less efficient than specialized algorithms when specialized algorithms are available, but inventing specialized algorithm is hard (both for us and for evolution) so often specialized algorithms simply aren't available. You don't seem to have responded to this line of argument..."

As far as I can tell you already answered that by saying that humans are a conglomerate of many specialized techniques rather than something that is similar to AIXI.

The confusion of which the question by Wei_Dai is a result is inherent in the vagueness of the concept of an agent.

Humans are not agents in the same sense that a cooperation, a country or humanity as a whole is not an agent.

What is usually referred to as an agent is an expected utility maximizer, a rigid consequentialist that acts in a way that is globally and timelessly instrumentally useful.

Which is not how effects of evolution work or how any computationally limited decision maker could possible work.

It takes complex values, motivations, a society of minds and its cultural evolution to yield behavior approximating agency.

Complex values are the cornerstone of diversity, which in turn enables creativity and drives the exploration of various conflicting routes. A singleton with a stable utility-function lacks the feedback provided by a society of minds and its cultural evolution.

You need to have various different agents with different utility-functions around to get the necessary diversity that can give rise to enough selection pressure. A “singleton” won’t be able to predict the actions of new and improved versions of itself by just running sandboxed simulations. Not just because of logical uncertainty but also because it is computationally intractable to predict the real-world payoff of changes to its decision procedures.

You need complex values to give rise to the necessary drives to function in a complex world. You can’t just tell an AI to protect itself. What would that even mean? What changes are illegitimate? What constitutes “self”? That are all unsolved problems that are just assumed to be solvable when talking about risks from AI.

An AI with simple values will simply lack the creativity, due to a lack of drives, to pursue the huge spectrum of research that a society of humans does pursue. Which will allow an AI to solve some well-defined narrow problems, but it will be unable to make use

of the broad range of synergetic effects of cultural evolution. Cultural evolution is a result of the interaction of a wide range of utility-functions.

2:57 AM

nick said...

humans are a conglomerate of many specialized techniques rather than something that is similar to AIXI.

A good thing, too, because "AIXI" (i.e. Solomonoff induction, a generalized model of machine learning) is uncomputable. Even learning with no guaruntees is harder than cracking public key crypto in the general case. To make learning efficient, you need to have some already existing information about the particular environment being learned, or similarly a learning algorithm that is specialized for that environment. The more relevant information you have, the better adapted the algorithm, the easier it is to learn the rest. Which gives us an economy of different learners specialized to different kinds of environments.

As for humans, we are certainly quite good at some important mental tasks, but there are also tasks we're relatively quite bad at. Exact memories typically being among them. Birds that can remember where in a landscape thousands of nuts are buried, the long-term memories of seals, animals with "photographic memories", etc. each in their own ways have specialized mental abilities our unaided brains don't typically have. And of course computers can memorize trillions of numbers, on top of being able to do arithmetic billions to trillions of times faster, and many other tasks that require boolean logic, arithmetic, and the like. And I'm not even counting the many sensory inputs our unaided brains lack, which before the advent of technological sensory modes caused us to miss out on most of the information available in our environment.

None of that stops humans from still having the best overall package of specialties, especially with the help of the technology we create, like those computers. But there is no magical "general intelligence" that we possess and other entities do not -- just a different package wrapped up in a brain with a higher brain/body ratio, hands for our brains to create that technology, and language that allows us to form some sophisticated social relationships that are very different from those of other animals. A very good bundle of specialties, but missing some pieces we'd really now love to have (like the memories of computers), and certainly not "general intelligence", which ranges from the astronomically inefficient to the uncomputable.

2:59 PM

nick said...

Speaking of comparative intelligence, here's a fun little debate about whether to use humans, monkeys, or robots to pick coconuts. Apparently monkeys can climb ten times as many trees in a day as humans, but they have a hard time telling whether a coconut is ripe. Not clear on what parts of the task the robots do better or worse:

http://ibnlive.in.com/news/coconut-plucking-is-monkey-business-in-kerala/253500-62-126.html

4:27 PM

Alexander Kruel said...

Wei Dai addressed your directly in his latest post at lesswrong.com, asking: *Work on Security Instead of Friendliness?*

Here is my reply:

For the sake of the argument I am going to assume that Nick Szabo shares the views of the Singularity Institute that at some point there will exist a vastly superhuman intelligence that cares to take over the world, which, as far as I can tell, he doesn't.

The question is not how to make systems perfectly secure but just secure enough as to disable such an entity from easily acquiring

massive resources. Resources it needs to create e.g. new computational substrates or build nanotech assemblers etc., everything that is necessary to take over the world.

In his post, Wei Dai claims that it is 2012 and that *"malware have little trouble with the latest operating systems and their defenses."*

This is true and yet not relevant. What can be done with modern malware is nowhere near sufficient to what an AI will have to do to take over the world.

An AI won't just have to be able to wreck havoc but take *specific* and *directed* control of highly volatile social and industrial institutions and processes.

It doesn't take anywhere near perfect security to detect and defend against such large scale intrusions.

12:12 PM

nick said...

Here was my reply:

(1) I'd rephrase the above [comment I posted under my "The Singularity" post on this blog, reposted to Less Wrong by Wei Dai] to say that computer security is among the two most important things one can study with regard to this alleged threat [the robot apocalypse, assuming one wanted to take it seriously].

(2) The other important thing is law. Law is the "offensive approach to the problem of security" in the sense I suspect you mean it (unless you mean something more like the military). Law is very highly evolved, the work of millions of people as smart or smarter than Yudkoswky over more than a millenium, and tested empirically against the real world of real agents with a real diversity of values every day. It's not something you can ever come close to competing with by a philosophy invented from scratch.

(3) I stand by my comment that "AGI" and "friendliness" are hopelessly anthropomorphic, infeasible, and/or vague.

(4) Computer "goals" are only usefully studied against actual algorithms, or clearly defined mathemetical classes of algorithms, not vague and imaginary concepts. Perhaps you can make some progress by for example advancing the study of postconditions, which seem to be the closest analog to goals in the software engineering world. One can imagine a world where postconditions are always checked, for example, and other software ignores the output of software that has violated one of its postconditions.

9:31 PM

Alexander Kruel said...

@Nick Szabo

Not sure if you saw the movie Eagle Eye, but it nicely highlights how those people probably perceive your law commentary.

You have to realize the following when arguing with those people.

Namely their premises, which are very roughly:

1.) There will soon be a superhuman intelligence.
2.) It will be a master of science and social engineering.
2.b.) It will be able to take over the world by earning money over the internet, control people by making phone calls and sending emails, and ordering equipment to invent molecular nanotechnology.
3.) It will interpret its goals in a completely verbatim way without wanting to refine those goals.

4.) It will want to protect those goals and in doing so take over the world and the universe.
5.) Any goal short of a full-fledged coherent extrapolation of the volition of all of humanity will lead to the destruction of human values either as a side-effect or because humans are perceived to be a security risk.

It all sounds incredible crazy, naive and absurd. But as long as you don't argue against those premises they will just conjecture arbitrary amounts of intelligence, i.e. magic, to "refute" any your more specific arguments.

12:15 PM

Alexander Kruel said...

On of the key pieces of evidence they like to cite is something they call the AI-Box Experiment. An unpublished, not reproducible chat between Eliezer Yudkowsky and a few strangers where he played the role of an AI trying to convince the gatekeeper, played by the other person, to release it into the world.

One of the rules, which basically allows the AI to assert anything, is as follows:

"The Gatekeeper must remain engaged with the AI and may not disengage by setting up demands which are impossible to simulate. For example, if the Gatekeeper says "Unless you give me a cure for cancer, I won't let you out" the AI can say: "Okay, here's a cure for cancer" and it will be assumed, within the test, that the AI has actually provided such a cure. Similarly, if the Gatekeeper says "I'd like to take a week to think this over," the AI party can say: "Okay. (Test skips ahead one week.) Hello again.""

12:27 PM

nick said...

Alexander, I hadn't quite realized just how far from reality they had traveled down their imaginary rabbit hole. I hope in the future I'll be finding more constructive things to do then trying to engage them. Still, I applaud you for your research of them and hopefully the result will be saving some people from such mind-wasting drivel.

5:42 PM

nick said...

This comment has been removed by a blog administrator.

10:37 PM

Dmytry said...

@nick: Wise choice.

With the AI box nonsense - why one would even bother building AI so general that it will want out of the box... I thought that this was about a boxed mind upload or other neural network derived AI.

Generally this singularity/rationalism community has very strong belief in powers of their reason (as per philosophy of rationalism), and they mix up rationalism and rationality.

When talking of the AGI that others allegedly would create, they point out how bad it would be and how it would kill mankind.

Whenever you point out to them that e.g. an algorithm which solves formally defined problems is very practically useful (and can be tackled directly without any further problems of defining what we want), they are trying to convince you that it is much less useful

than their notion of AGI that wants real things, having forgotten the previous statement, or being unable to see how generally useful something like mathematics is.

It's true 1984 doublethink over there. Almost everyone with any sense has evaporated, while the very few people with a clue (Wei Dai for example) read their own sense into the scriptures, arguing something sensible but entirely orthogonal to the nonsense of the whole group.

12:20 PM

BenW said...

Eventually the unexpected will happen.
But that does not mean every hypothesis will come true.
Principles of black swan events and being anti fragile in our preparations still seem to apply. Having a first aid kit in your house may never need to be used. But if a scenario arises where it does, that $40 investment could save a life.

8:44 AM

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)

Simple theme. Powered by Blogger.